arXiv:2305.01980v1 [cs.SD] 3 May 2023

DIVER 基于文本描述生成多样且生动的声音

Guangwei Li、*Xuenan Xu*、*Lingfeng Dai*、*Mengyue Wu*†、*Kai Yu*†

X-LANCE实验室，计算机科学与工程系，人工智能教育部重点实验室，人工智能研究院，上海交通大学，中国上海

摘要

以往的音频生成主要聚焦于语音或音乐等特定声音类别，其形式和内容受到很大限制。本文通过使用自然语言描述作为线索来生成广泛的声音，突破了特定音频生成的范畴。不同于视觉信息，文本描述本质上较为简洁，但其背后蕴含着丰富的隐藏含义，这给待生成音频带来了更高的可能性和复杂性。我们采用变分量化GAN进行训练，学习一个用于表示频谱图离散表征的码本。对于给定的文本描述，我们将其预训练嵌入输入Transformer，对码本索引进行采样，从而解码出频谱图；随后通过MelGAN声码器将其进一步转换为波形。生成的波形具有高质量和高保真度，同时与给定文本高度对应。实验表明，我们提出的方法能够生成自然、生动的音频，并取得出色的定量和定性结果。

索引词— 声音生成、变分量化 GAN、文本到声音

1. 引言

跨模态生成任务受到了越来越多的关注，涵盖从文本到图像 [1, 2, 3]，图像到音频 [4, 5, 6],以及反向生成。然而，基于文本生成音频，尤其是根据句子级描述进行音频生成，仍处于起步阶段。由于现实世界中声音事件存在时频相关性，音频生成面临着独特的挑战。例如，对于句子“食物正在煎炸，一位女性正在说话”，两种声音在时间先后顺序和持续时长上的不确定性，会极大地影响可能生成的音频内容。在本文中，我们定义了一项任务，即根据复杂音频生成自由文本描述（见图1）。生成的音频应保持高质量和自然度，并全面反映句子中描述的声音类别及其相互关系。

以往的音频生成任务主要聚焦于合成特定的声音类别，例如语音 [7, 8, 9]和音乐 [10, 11, 12]。然而，这些任务还可以进一步扩展：i）待生成的声音类别是预先定义的，因此输入受到限制；ii）生成的内容实际上是对输入模态的转换。一项新兴研究 [13] 聚焦于以文本为输入进行音频生成，利用拟声词合成环境声音。然而，由于拟声词有限，生成内容的多样性仍然受到限制

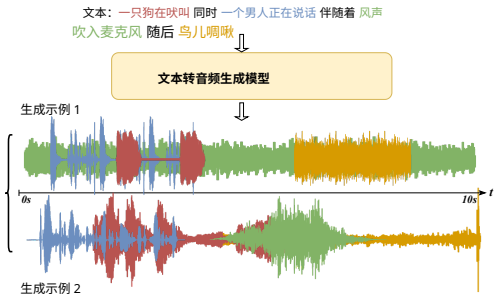

图 1。所提出的基于野外文本的音频生成旨在根据给定的任意文本生成高质量、内容丰富的音频。这里展示了模型根据一个包含多个声音事件的句子生成的两个示意性示例。

有时表现力也较弱。最近，研究者借助视觉线索 [4, 5, 6]或拟声词 [13]，实现了输入限制较少的生动音频生成。然而，基于自由文本描述生成多样化音频的研究仍然很少。专注于这一任务的同期研究包括 [14, 15]。与可从原始视频来源自然获取的音频-视频或音频-图像对不同，音频-文本对需要额外标注。对同一段音频的描述性标注在用词、整体风格和行文顺序上可能差异很大。

受视觉线索辅助音频生成研究的启发，我们提出了一种文本到音频生成系统，能够根据描述性字幕生成高质量的音频。该系统如图2所示。Transformer 和参考码本根据预训练语言模型提取的文本特征生成稠密表示。随后，解码器将码本表示转换为频谱图。最后，频谱图通过 MelGAN [16]声码器转换为波形。码本和解码器均来自预训练的变分量化生成对抗网络（VQGAN） [4, 1]。该系统使用音频描述数据集 [17, 18], 进行训练，其中提供了音频片段及其对应的文本描述。

我们采用包括 PSNR、FID、MMKL 和 SPICE、CIDEr（详见第 3.1 节）在内的多种指标，以及主观人工分析进行评估。定量和定性结果均表明，生成音频具有较高的听觉质量，并能反映给定的文本，从而体现了我们文本到音频生成方法的有效性。此外，我们还展示了一些生成示例，以便更直观地进行呈现[1]。

† Mengyue Wu 和 Kai Yu 为通讯作者。本研究得到上海市科技重大项目（2021SHZDZX0102）和江苏省科技项目（编号：BE2022059-2）的资助。

[1] https://ligw1998.github.io/audiogeneration.html

©2023 IEEE。允许将本材料用于个人用途。对于其他所有用途，包括在任何当前或未来的媒体中重印/再版本材料、将本材料用于广告或宣传目的、创作新的汇编作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分，必须获得 IEEE 的许可。

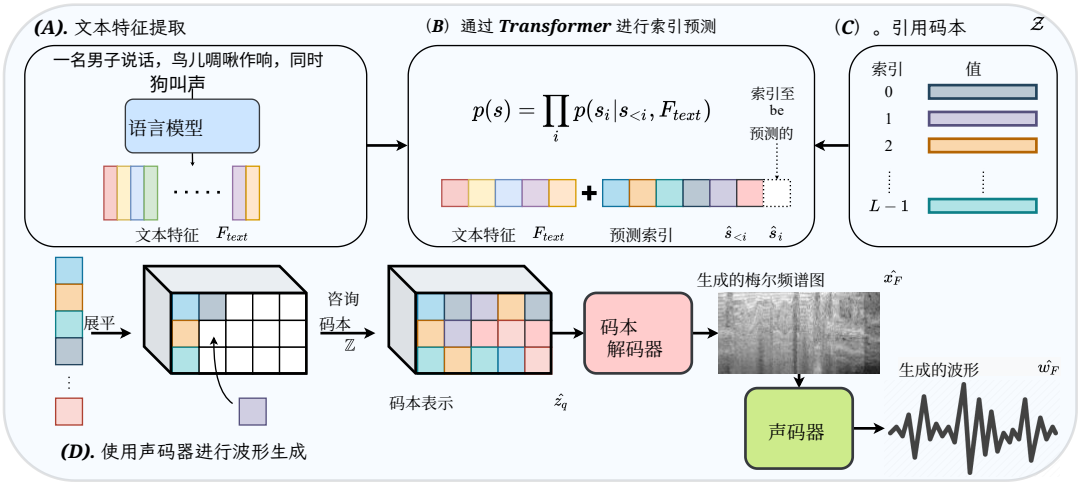

图 2。我们文本到音频生成系统的主要结构。采样时包含四个主要步骤。在 (A) 中，通过预训练语言模型从给定句子中提取文本特征。在 (B) 中，根据文本特征自回归地生成码本索引。(C) 使用预训练码本，根据 (B) 中生成的索引获取码本表示。随后，表示被转换为梅尔频谱图，再由 (D) 中的声码器将其转换为波形。

2. 文本到音频生成

我们的主要目标是生成能够全面反映给定文本描述中的可听元素，同时保持高质量和高保真度的音频。以下各节将相应介绍主要部分，框架概览如图2所示。

2.1. 学习到的表示的离散频谱图码本

如上所述，我们的目标首先是根据输入文本构建音频梅尔频谱图。一种直接的方法是预测频谱图中每个像素的最优值。然而，原始频谱图中的像素数量可能极其庞大，使得 Transformer 难以进行注意力计算。

为了更高效地表示梅尔频谱图，我们使用离散码本来表示频谱图。我们的方法借鉴 VQGAN 来训练码本，该方法最初发表于 [1]，并在 [4] 中应用于音频。VQGAN 是一种自编码器，其表示尺寸小于向量量化变分自编码器（VQVAE），[19]最初用于解码图像。在此，我们采用 VQGAN 重建给定的频谱图 $x \in R^{M\times T}$，尺寸为 $\hat{x} \in R^{M\times T}$，$M$ 表示梅尔滤波器组的数量，而 T 表示时间维度。

在训练过程中，频谱图通过编码器 E 编码为一个较小的 $\hat{z} \in \mathbb{R}^{m\times t\times n_z}$，其中 n_z 充当编码维度。随后，可以将这一小规模表示展平为一个 $m \times t$ 序列，以确定从学习得到的码本 $\mathcal{Z} = \{z_l\}_{l=1}^{L} \subset \mathbb{R}^{n_z}$ 中选取的条目。我们采用逐元素量化 q，将 $\hat{z}$ 中的每个独立 $\hat{z}_{ij}$ 映射到与其最接近的码本条目 z_l，具体可描述为：

$$z_q = q(\hat{z}) := \arg\min_{z_l\in\mathcal{Z}} \|\hat{z}_{ij} - z_l\| \tag{1}$$

z_q 是量化表示，可通过解码器 D进一步解码为频谱图$\hat{x}$:

$$\hat{x} = D(z_q) = D(q(\hat{z})) \tag{2}$$

我们的目标是尽可能减小 x 与 an 之间的差异 d$\hat{x}$.

尽管公式 (1) 中的量化过程不可微分，但根据 [20], ，公式 (2) 中用于反向传播的梯度可以通过梯度估计器从编码器复制到解码器。公式 (2) 的损失函数可表示为：

$$\mathcal{L}_{VQ}(E, D, \mathcal{Z}) = \|\hat{x} - x\|^2 + \|\text{sg}[\hat{z}] - \text{z}_\text{q}\|_2^2 + \beta\|\text{sg}[\text{z}_\text{q}] - \hat{\text{z}}\|_2^2 \tag{3}$$

其中，$\mathcal{L}_{rec} = \|\hat{x} - x\|^2$ 和 sg[·] 表示停止梯度运算，该运算在反向传播期间的梯度为零，而 β 是一个加权因子。

VQGAN 在公式（3）中的损失函数基础上，增加了判别器 [21] 和感知损失 [22]。我们遵循 [4]，并针对频谱图对 VQGAN 进行调整，同时将原始感知损失替换为学习感知音频块相似度。Spectrogram VQGAN 的最终损失函数为：

$$\mathcal{L}_{SpecVQGAN}(\{E, D, \mathcal{Z}\}, C) = \mathcal{L}_{VQ}(E, D, \mathcal{Z}) + logC(x) + log(1 - C(\hat{x})) + \sum_s \frac{1}{M^s T^s}\|x^s - \hat{x}^s\|_2^2 \tag{4}$$

这里，C 是一个基于图块的判别器，x^s 和$\hat{x}^s$ 是来自在 AudioSet 上训练的 VGGish-ish [4] 的 s^{th} 尺度的真实频谱图和重建频谱图的特征。

2.2. 文本特征提取

在训练 Transformer 对码本索引进行采样之前，需要提取文本描述，以提供生成音频片段所需的信息。句子被输入基于**BERT**的 [23] 模型，以转换为嵌入序列。我们还使用**CLIP** [24]的文本编码器。受 CLIP 启发，我们还采用对比学习范式进行音频-文本预训练。预训练后，文本编码器被用作高级特征提取器，称为**CLAP** [25]。与 CLIP 类似，在每个批次中，我们将匹配的音频-文本对作为正样本，将不匹配的音频-文本对作为负样本。我们在一个整合数据集上训练模型，该数据集基于目前可获得的公开文本-音频数据集（AudioCaps、Clotho、MACS）。

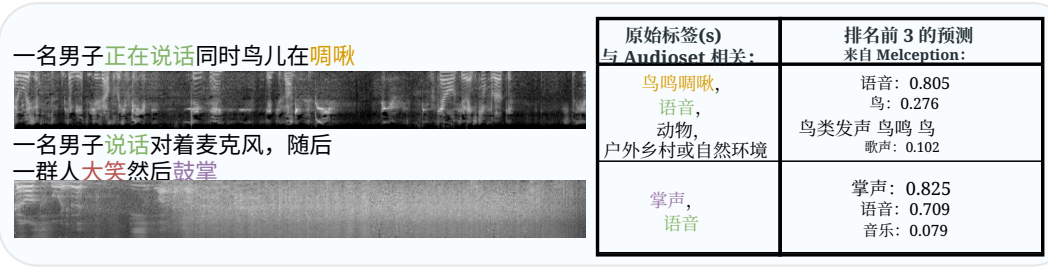

图 3。文本转音频生成系统的示例。我们将生成的频谱图输入 Melception 音频分类器，右侧列出了事件检测结果，以及 AudioSet 中原始音频对应的真实标签。

2.3. 使用 Transformer 进行文本条件生成

通过编码器 E 和解码器 D （在“离散化”章节中训练得到），以及在第2.2节中提取的文本特征，我们能够通过潜在 Transformer 生成频谱图。给定文本特征 F_{text}，Transformer 学习预测下一个索引的分布s_i，其中的索引为 $s_{<i}$。序列 $p(s)$ 的似然为：

$$p(s|F_{text}) = \prod_i p(s_i|s_{<i}, F_{text}) \tag{5}$$

Transformer 通过最大化条件序列的对数似然进行训练：

$$\mathcal{L}_{Transformer} = \mathbb{E}_{x\sim p(x)}[-logp(s|F_{text})] \tag{6}$$

推理时，码本索引 $\hat{}s_i$ 从 p 提供的分布中采样。自回归生成，直到 $i = m \times t$，索引被展平为形状 (m, t)，并根据 [4] 以列优先的方式排列；随后用每个码在码本中的索引替换它，并据此查找码本 $\mathcal{Z}$：

$$(z_q)_{ij} = z_l \text{ if } s_{ij} = l \tag{7}$$

其中 $i \in \{1, 2, 3, ..., m\}$ 和 $j \in \{1, 2, 3, ..., t\}$。通过将索引映射回码本表示 $\hat{}z_q \in \mathbb{R}^{m\times t\times n_z}$，可以使用解码器 D恢复生成的频谱图：

$$\hat{x_F} = D(\hat{z_q}) \tag{8}$$

3. 实验

3.1. 数据集和评估指标

我们在两个数据集上进行实验：AudioCaps [18] 和 Clotho [17], , 并从客观和主观两个方面评估性能。

客观评估 我们采用多种指标评估生成性能。使用 Fréchet起始距离 [26], 、峰值信噪比和 MMKL 衡量生成频谱图的质量。此外，文本生成的高质量音频应能够通过音频描述模型 [25]反向生成描述。我们根据 SPICE [27] 和 CIDEr [28] 计算输入文本与反向生成文本之间的相似度，以衡量输入文本与生成音频之间的相关性。

多类别 Melception KL 散度（MMKL）是 MKL 的改进版本，最初在 [4] 中提出。原始分类器 Melception 使用 VGGSound 进行训练，每个片段仅对应一个声音标签；而 AudioCaps 中的大多数音频片段包含两个以上的声音类别，因此我们将 Softmax 激活层替换为 Sigmoid，并使用二元交叉熵损失在 AudioSet 上训练 Melception，以进行评估。

主观评估我们对生成的波形的质量和相关性均进行了人工评估。对于给定的每条文本描述，四名评分员分别从相关性和生成音频的质量两个方面进行评分，评分范围为 0—10 分。具体而言，“相关性”衡量音频与文本描述的匹配程度，“质量”则表示生成的声音与现实生活中的声音相似的程度，以及辨认这些声音的难易程度。

3.2. 实验设置

在数据预处理中，每个音频片段的时长为 ≈10 秒，采样率为 32000Hz，并以 12.5ms 的间隔进行提取，窗口大小为 50ms。梅尔滤波器组的数量为 80。我们采用了不同的预训练语言模型，包括 BERT-base-uncased、BERT-medium、CLIP 和 CLAP 文本编码器，将文本描述转换为取自隐藏状态的特征。对于所有文本编码器，最大序列长度均设置为 30。

在训练码本时，我们将码本的长度 $|\mathcal{Z}|$ 设为 1024。为稳定模型，在码本训练过程的初始阶段，将公式（4）中的损失对抗部分置零。对于 Transformer 部分，我们采用基于 GPT-2 的架构 [29], , 隐藏维度为 1024。Transformer 的输出经过一个以 Softmax 结尾的 1024 类分类器，从而学习下一个码本索引的分布。在推理过程中，我们将下一个码本标记的分布截断为 Top-K 概率，以控制样本的多样性。K 的范围从 K = $|\mathcal{Z}|$ = 1024 到 K = 1，此处我们选择 K=512。用于 MMKL 的 Melception 在 AudioSet 上进行训练，mAP 达到 33.4%。我们采用基于 MelGAN 的架构，以弥合预测的梅尔频谱图与波形之间的差距。

4. 结果

在本节中，我们根据客观指标和主观指标，对 AudioCaps 和 Clotho 数据集上生成的音频样本进行评估。

4.1. 定量结果

AudioCaps 我们的最终结果通过评估生成的频谱图与真实频谱图之间的相似度计算得出，如表 1 所示。由于此前尚无关于文本到自然音频生成的研究，我们将最佳模型（h）与第 2.2 节中提及的采用不同文本编码器的模型进行比较。此外，鉴于先前的研究表明降维特征的有效性和实用性，我们还调整了输入模型的特征序列长度 [23]。目前对特征长度的比较包括 无特征 （0个特征，即用随机嵌入替换文本特征）、1个特征 （1个特征

表1。不同配置下系统的性能

条件	文本编码器	峰值信噪比↑	FID↓	MMKL↓	SPICE↑	CIDEr↑
(a) 无特征	—	13.69	0.977	15.80	0.035	0.053
(b) 1个特征	BERT-medium	14.62	0.931	10.98	0.073	0.180
(c) 1个特征	BERT-base	14.78	0.936	10.43	0.075	0.183
(d) 1个特征	CLIP	14.89	0.918	10.00	0.083	0.194
(e) 1个特征	CLAP	**14.97**	0.784	8.74	0.092	0.246
(f) 30个特征	BERT-medium	14.66	0.858	10.06	0.081	0.190
(g) 30个特征	BERT-base	14.92	0.851	10.05	0.085	0.210
（h）30 个特征	CLAP	14.83	**0.778**	**8.57**	**0.096**	**0.249**

来自 [CLS] 标记的输出嵌入（第一个隐藏状态的输出嵌入）和 30 个特征（完整输出特征）。（a）列出了没有任何输入条件时的结果。我们将其设定为 MMKL 的上界；（b、c、d、e）比较了在 1 个特征条件下不同文本特征提取器的影响；类似地，（f、g、h）重点研究了使用不同特征尺寸时完整 30 个特征的表现。

可以从结果中得出几点结论：1）在不输入文本条件的情况下进行采样时，如图 (a) 所示，PSNR 和 MMKL 的上界分别为 13.69 和 15.80，而字幕生成指标相对较低；2）至少添加一个文本特征后，这两个指标将显著变化，分别达到 14。78↑ 和 10。43↓ 如图 (c) 所示，字幕生成结果也有所提升，这表明输入文本信息对生成音频至关重要；3）增加文本特征的数量将提高相关性（MMKL 更低），并提升整体生成质量（Fréchet起始距离更低）（f、g、h）。因此我们推断，输入信息越多，模型就越能区分输入句子中的声源及其关系。4）使用从较小语言模型中提取的、特征维度较小的文本特征，会对生成质量产生轻微的负面影响（d、e）（b、c）。不过，这并非关键因素，因为根据句子生成音频内容时，重点在于句子中提及的声音类别及其发生时序关系。5）我们的最佳模型是采用 CLAP 的 (h)，这意味着通过音频-文本对比学习提取的特征，更善于表示给定原始文本中的音频事件。图 3 展示了可视化的样本频谱图及 Melception 的音频标注结果3。

Clotho总体而言，如表 2所示，模型在 Clotho 上的音频生成性能不如在 AudioCaps 上的表现，因为 Clotho 的数据集规模明显小于 AudioCaps，且 Clotho 中的文本描述更具文学性。与 AudioCaps 相比，Clotho 使用了更复杂的句法和术语，这使模型更难解析文本，并建立文本与音频之间的联系。如果使用在 AudioCaps 上训练的第一阶段码本，并在 Clotho 上训练 Transformer，结果会有所改善，这表明在 AudioCaps 上训练出的码本更强大、更有效。

表 2。在 Clotho 测试集上的结果，码本使用两个数据集进行训练。

码本	MMKL↓	SPICE↑	CIDEr↑
Clotho	11.613	0.049	0.091
AudioCaps	10.988	0.068	0.137

4.2. 定性分析

我们从 AudioCaps 测试集中生成的 964 个样本中随机选取了 200 个。对于每条文本描述，我们向评审者提供四段音频：AudioCaps 中的原始音频；未提供文本特征时生成的音频；以一个特征为条件生成的音频；以及以完整特征为条件生成的音频。平均结果列于表 3。

表 3。人工评估结果。生成的音频片段分别从相关性和质量两个方面进行评分，评分范围为 0～10 分。

条件	平均相关性（/10）	平均质量（/10）
原始	**9.55**	**9.55**
无特征	1.85	4.58
BERT-b	5.50	6.08
CLAP	**6.79**	**6.13**

可以观察到，生成音频的相关性和质量都与所使用的文本特征相关。使用音频预训练的 CLAP 提取器时，生成样本的听觉反馈优于使用其他提取器得到的结果，这与上一节列出的定量结果一致。然而，生成音频与原始音频片段之间仍存在显著差距。咨询评审人员后，我们总结出当前生成音频的一些局限：1）文本描述中提及的不同声音类别可能交替出现，而这在现实生活中并不常见；2）单个片段中不同声音的音量可能相差很大；3）某些罕见的声音类别有时会被遗漏；4）生成的“语音”内容含糊不清，没有确切的词语。

5. 结论

本文聚焦于一项新颖且富有挑战性的任务：根据文本描述生成高质量且多样化的声音内容。借助由 VQGAN 训练的参考码本，我们的模型能够从更小的表示中高效地重建梅尔频谱图。采样时，我们采用基于 Transformer 的结构来弥合输入文本特征与预测表示之间的差距。我们的最佳模型在定量指标峰值信噪比（PSNR）、Fréchet起始距离（FID）和 MMKL 上分别取得了 14.83、0.778 和 8.57 的平均得分，同时在音频描述结果上，SPICE 得分为 0.096，CIDEr 得分为 0.249。人工评估也证实了生成音频片段的整体质量，以及其与给定文本的相关性。这两项结果均表明，我们提出的方法能够生成与文本所传达信息相对应的自然且多样化的音频。我们还期待未来开展进一步研究，包括改善生成声音事件的时间顺序和持续时长。

6. 参考文献

[1] Patrick Esser、Robin Rombach 和 Bjorn Ommer，《驯服用于高分辨率图像合成的变换器》，载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第 12873–12883 页。[2] Alex Nichol 等 Dhariwal，《Glide：迈向具有照片真实感的图像生成与编辑——文本引导扩散模型》，*arXiv* 预印本 *arXiv:2112.10741*，2021年。[3] Chitwan Saharia、William Chan、Saurabh Saxena 等 Li，《具有深度语言理解的照片真实感文本到图像扩散模型》，*arXiv* 预印本 *arXiv:2205 11487*，2022。[4] Vladimir Iashin 和 Esa Rahtu，《驯服视觉引导的声音生成》，*arXiv* 预印本 *arXiv:2110.08791*，2021年。[5] Kun Su 等，《Audeo：为无声表演视频生成音频》，神经信息处理系统进展，第 33 卷，第 3325–3337 页，2020年。[6] Huadong Tan、Guang Wu、Pengcheng Zhao 和 Yanxiang Chen，《通过自注意力进行声谱图分析以实现跨模态视听生成》，载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020年，第 4392–4396 页。[7] Jonathan Shen、Ruoming Pang、Ron J Weiss、Rj Schuster 等，《通过以梅尔频谱图预测为条件的 WaveNet 实现自然的文本转语音合成》，载于 *2018*年 *IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第 4779–4783 页。[8] Naihan Li、Shujie Liu、Yanqing Liu、Sheng Zhao、Ming Liu 和 Ming Zhou，《借助 Transformer 实现接近人类质量的文本转语音合成》，*arXiv* 预印本 *arXiv:1809.08895*，2018年。[9] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速且高质量的端到端文本转语音》，*arXiv* 预印本 *arXiv:2006.04558*，2020年。[10] Li-Chia Yang 等 Chou，《MidiNet：用于符号域音乐生成的卷积生成对抗网络》，*arXiv* 预印本 *arXiv:1703.10847*，2017年。[11] Hao-Wen Dong、Wen-Yi Hsiao、Li-Chia Yang 和 Yi-Hsuan Yang，《MuseGAN：用于符号音乐生成与伴奏的多轨顺序生成对抗网络》，载于 *AAAI*人工智能会议论文集，2018年，第 32 卷。[12] Bryan Wang 和 Yi-Hsuan Yang，《PerformanceNet：利用多频带卷积残差网络进行从乐谱到音频的音乐生成》，载于*AAAI*人工智能会议论文集，2019年，第 33 卷，第 1174–1181 页。[13] Yuki Okamoto 等 Imoto，《 Onoma-to-wave：从拟声词合成环境声音》，*arXiv* 预印本 *arXiv:2102.05872*，2021年。[14] Dongchao Yang、Jianwei Yu、Helin Wang、Wen Wang、Chao Weng、Yuexian Zou 和 Dong Yu，《DiffSound：用于文本到声音生成的离散扩散模型》，*arXiv* 预印本 *arXiv:2207.09983*，2022年。[15] Felix Kreuk、Gabriel Synnaeve、Adam Polyak、Uriel Singer、Alexandre D´efossez、Jade Copet、Devi Parikh、Yaniv Taigman 和 Yossi Adi，《AudioGen：文本引导的音频生成》，*arXiv* 预印本 *arXiv:2209.15352*，2022年。

[16] Kundan Kumar、Rithesh Kumar、de Boissiere 等，《Mel-gan：用于条件波形合成的生成对抗网络》，神经信息处理系统进展，第 32 卷，2019年。[17] Konstantinos Drossos、Samuel Lipping 和 Tuomas Virtanen，《Clotho：一个音频描述数据集》，载于*ICASSP 2020-2020 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第 736–740 页。[18] Chris Dongjoo Kim、Byeongchang Kim、Hyunmin Lee 和 Gunhee Kim，《Audiocaps：为真实场景中的音频生成描述》，载于*2019*年北美计算语言学协会年会论文集：人类语言技术，第 *1* 卷（长篇与短篇论文），2019年，第 119–132 页。[19] Aaron Van Den Oord、Oriol Vinyals 等，《神经离散表示学习》，神经信息处理系统进展，第 30 卷，2017年。[20] Yoshua Bengio 和 L´eonard 等，《针对条件计算，通过随机神经元估计或传播梯度》，*arXiv* 预印本 *arXiv:1308.3432*，2013年。[21] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A Efros，《使用条件对抗网络进行图像到图像转换》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第 1125–1134 页。[22] Justin Johnson、Alexandre Alahi 和 Li Fei-Fei，《用于实时风格迁移与超分辨率的感知损失》，载于欧洲计算机视觉会议。Springer，2016年，第 694–711 页。[23] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，*arXiv* 预印本 *arXiv:1810.04805*，2018年。[24] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，《通过自然语言监督学习可迁移的视觉模型》，载于国际机器学习会议。PMLR，2021年，第 8748–8763 页。[25] Xuenan Xu、Zeyu Xie、Mengyue Wu 和 Kai Yu，《上海交通大学 DCASE2022 挑战赛任务 6 系统：采用音频文本检索预训练的音频描述》，技术报告，DCASE2022 挑战赛，2022年。[26] Martin Heusel、Hubert Ramsauer、Unterthiner 等，《通过双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡》，神经信息处理系统进展，第 30 卷，2017年。[27] Peter Anderson、Basura Fernando、Mark Johnson 和 Stephen Gould，《SPICE：基于语义命题的图像描述评估》，载于欧洲计算机视觉会议。Springer，2016年，第 382–398 页。[28] Ramakrishna Vedantam、C Lawrence Zitnick 和 Devi Parikh，《CIDEr：基于共识的图像描述评估》，载于*IEEE*计算机视觉与模式识别会议论文集，2015年，第 4566–4575 页。[29]Klemens Lagler、Michael Schindelegger、Johannes B¨ohm、Hana Kr´asn´a 和 Tobias Nilsson，《GPT-2：无线电空间大地测量技术的经验倾斜延迟模型》，《地球物理研究快报》，第 40 卷，第 6 期，第 1069–1073 页，2013年。